(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 March 2004 (04.03.2004)

PCT

(10) International Publication Number
WO 2004/019230 A3

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number:
PCT/US2003/026025

(22) International Filing Date: 20 August 2003 (20.08.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/404,581 20 August 2002 (20.08.2002) US
10/293,859 13 November 2002 (13.11.2002) US

(71) Applicant *(for all designated States except US)*: MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD. [JP/JP]; Matsushita IMP Bldg., 19F, 1-3-7, Shiromi, Shuo-ku, Osaka 540-6319 (JP).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: GUO, Jinhong, Katherine [US/US]; 6 Tiffany Court, West Windsor, NJ 08550 (US). MA, Yue [US/US]; 6 Tiffany Court, West Windsor, NJ 08550 (US).

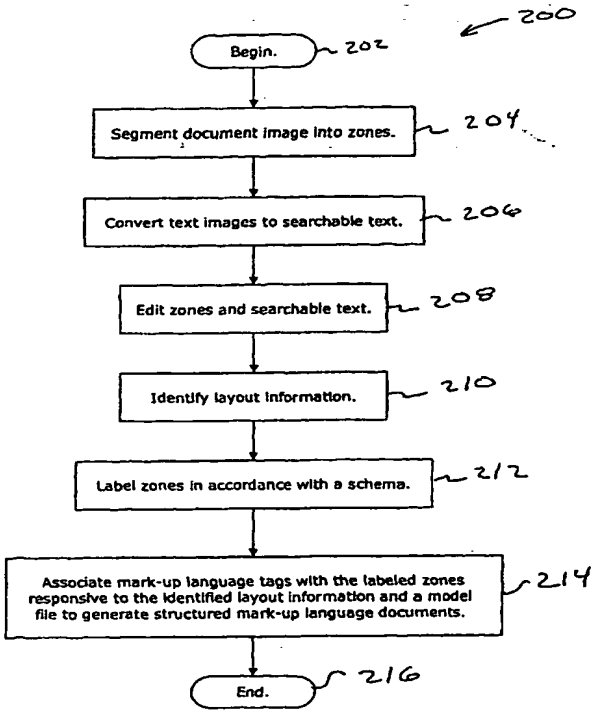
(74) Agent: NIGON, Kenneth, N.; RatnerPrestia, P.O. Box 980, Valley Forge, PA 19482 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: METHOD, SYSTEM, AND APPARATUS FOR GENERATING STRUCTURED DOCUMENT FILES



(57) Abstract: A method, system, apparatus, and graphical user interface (GUI) for generating structured document files from a document image is disclosed. Structured document files are generated by segmenting the document image into one or more zones containing respective text images, converting the respective text images to digital text, automatically identifying layout information for each of the one or more zones, labeling each of the one or more zones in accordance with a schema, and automatically associating mark-up language tags with the labeled zones to generate the structured document files responsive to the identified layout information and a model file.

WO 2004/019230 A3

Declaration under Rule 4.17:
— *of inventorship (Rule 4.17(iv)) for US only*

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
25 March 2004

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

Interna l Application No
PCT/US 03/26025

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	VALVENY E ET AL: "SCAN-TO-XML: AUTOMATIC GENERATION OF BROWSABLE TECHNICAL DOCUMENTS" , PROCEEDINGS OF THE INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, XX, XX, VOL. 3, PAGE(S) 188-191 XP001151841 page 189, paragraph 2 -page 190, paragraph 3 --- -/--	1-20

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
6 February 2004	18/02/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax (+31-70) 340-3016	May, M

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Internat	Application No
	PCT/US 03/26025

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
X	B.LAMIROY, L.NAJMAN: "Scan-to-XML: Using Software Component Algebra for Intelligent Document Generation" LECTURE NOTES ON COMPUTER SCIENCE, GRAPHICS RECOGNITION. ALGORITHMS AND APPLICATIONS. 4TH INTERNATIONAL WORKSHOP, GREC 2001, vol. 2390, 7 - 8 September 2001, XP009025490 Kingston, Ont., Canada paragraph '0003!; figure 1	1-20
Y	EP 0 854 433 A (MATSUSHITA ELECTRIC IND CO LTD) 22 July 1998 (1998-07-22) the whole document	1-20
Y	WO 00 56033 A (ORACLE CORP) 21 September 2000 (2000-09-21) the whole document	1-20
Y	US 6 327 388 B1 (LOPRESTI DANIEL P ET AL) 4 December 2001 (2001-12-04) the whole document	1-20
Y	INTERNATIONAL BUSINESS MACHINES CORPORATION: "Conversion of final form data, such as AFP, to XML" RESEARCH DISCLOSURE, KENNETH MASON PUBLICATIONS, HAMPSHIRE, GB, vol. 444, no. 208, April 2001 (2001-04), XP007128106 ISSN: 0374-4353 the whole document	1-20
Y	INTERNATIONAL BUSINESS MACHINES CORPORATION: "Conversion of style based documents to arbitary XML formats using externalized rule database" RESEARCH DISCLOSURE, KENNETH MASON PUBLICATIONS, HAMPSHIRE, GB, vol. 460, no. 120, August 2002 (2002-08), XP007131058 ISSN: 0374-4353 the whole document	1-20

INTERNATIONAL SEARCH REPORT

...rmation on patent family members

Interna... Application No
PCT/US 03/26025

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
EP 0854433 A	22-07-1998	US 5892843 A	06-04-1999
		DE 69724755 D1	16-10-2003
		EP 0854433 A2	22-07-1998
		JP 10260993 A	29-09-1998
WO 0056033 A	21-09-2000	AU 759477 B2	17-04-2003
		AU 3748300 A	04-10-2000
		CA 2368089 A1	21-09-2000
		EP 1166524 A1	02-01-2002
		JP 2002539547 T	19-11-2002
		WO 0056033 A1	21-09-2000
US 6327388 B1	04-12-2001	NONE	